第35卷 第17期 计 算 机 工 程 2009年9月
Vol.35 No.17 Computer Engineering September 2009
·开发研究与设计技术· 文章编号：1000—3428(2009)17—0235—03 文献标识码：A 中图分类号：TP302.1

中国城市空气质量数据仓库的研究与实现

纪翠玲[1]，肖永康[2]
(1. 中国气象局培训中心，北京 100081；2. 北京师范大学信息科学与技术学院，北京 100875)

摘 要：根据空气质量监测数据的空间特征，对数据仓库的星型模型进行扩展，利用Oracle10g 和ArcGIS9.0设计并建立中国城市空气质量数据仓库的原型系统，实现面向高级分析、挖掘和决策的数据组织和管理。为方便查询和分析数据，设计一个基于Web 的空间OLAP系统，弥补了传统OLAP在空间分析和可视化表达方面的不足。

关键词：空气质量；数据仓库；空间OLAP

Research and Implementation of Chinese City Air Quality Data Warehouse

JI Cui-ling[1], XIAO Yong-kang[2]
(1. Training Center, China Meteorological Administration, Beijing 100081;
2. College of Information Science and Technology, Beijing Normal University, Beijing 100875)

【Abstract】This paper extends the star model according to the spatial characteristic of air quality monitor data, and utilizes Oracle10g and ArcGIS 9.0 to design and implement an air quality data warehouse prototype system in China. The system can organize and manage the data effectively to support the requirement of complicated analysis, mining and decision-making. To query and analyze the data in the data warehouse conveniently, this paper designs a Web spatial OLAP system to improve the ability of spatial analysis and visual presentation of the traditional OLAP system.

【Key words】air quality; data warehouse; spatial OLAP

1 概述

空气质量状况是全社会共同关注的问题。为掌握空气质量状况和污染变化趋势，我国在近200个城市建立了自动监测系统，全天候地对SO_2, NO_2和可吸入颗粒物等主要污染物的浓度值进行监测。经过多年监测，大部分监测站点已经积累了大量的历史数据[1-2]。有效管理和利用这些数据资源，对于研究中国空气质量状况的时空变化特征、指导空气污染防治决策、产业规划和布局具有重要意义。目前，这些监测数据主要存储在文本文件、Excel 表格、SQL Server 或Oracle数据库中，并按照监测时间或监测点进行管理和组织。这种基于业务的数据管理方式和以此为基础的数据查询系统只能提供简单的查询功能，远远不能满足研究人员和决策者进行多尺度、多角度的复杂查询和综合分析的需求。

数据仓库是20世纪90年代发展起来的一种体系化的数据存储环境。与事务型数据库相比，数据仓库是面向主题的、集成的、时变的数据集合，支持强大的管理决策过程，能够有效解决海量的、历史时态数据的存储、管理和多维分析等问题[3]。但数据仓库技术擅长操纵关系型的数值型数据，目前还无法实现对空间数据的查询、分析和展现。

本文以86个城市2000年~2007年的日监测数据为例，利用Oracle10g和ArcGIS9.0建立了空气质量数据仓库的原型系统。为方便用户使用数据仓库中的数据，本文综合利用Web, OLAP和GIS 3种技术建立了Web Spatial OLAP系统。

2 空气质量数据仓库设计

2.1 空气质量数据仓库的体系结构

空气质量数据的获取、处理和汇交过程具有很强的自下而上的特点，即各县(区)的监测站将实时监测数据，汇集到市环保局，再向上级部门(如省环保局)汇交，最终汇集到中国环境监测总站，由国家环保总局负责全国数据的管理和对外发布[1-2]。根据这些特点，本文采用自底向上的模式设计中国城市空气质量数据仓库，如图1所示。

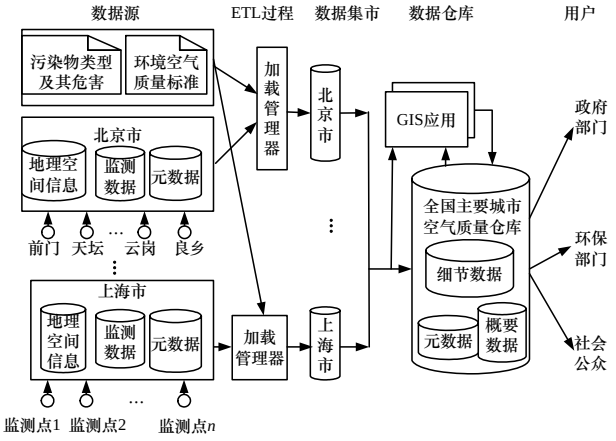

图1 中国城市空气质量数据仓库的体系结构

该模式的核心是各主要城市的环境保护部门首先建立数据集市，通过ETL过程将各自数据源中的数据进行清洗、转换，加载到数据集市中，然后将各数据集市中的数据逐步汇

基金项目：国家自然科学基金资助项目(60603068)
作者简介：纪翠玲(1979－)，女，助理研究员、博士，主研方向：地理信息共享，数据仓库；肖永康，讲师、博士
收稿日期：2009-02-17 **E-mail：**xiaoyk@bnu.edu.cn

交和集成，最终建立全国性的数据仓库。中央仓库中存储不同粒度级别的数据，以供政府部门、环境保护部门或社会公众等用户访问。在建立数据集市和数据仓库时，本文使用 ArcGIS9.0 管理数据源中的地图数据(如行政区图和监测点空间分布图等)，以便实现空间数据的查询、分析和展现等功能。使用自底向上模式的优点是，能够以最少的投资满足各城市的当前需求，然后再逐步扩充和完善，建立起国家级、省级和市级的 3 层管理模式，以支持不同级别部门的空气质量环境评价、管理和决策。

2.2 空气质量数据仓库的逻辑设计

逻辑设计主要考察数据库对象间的逻辑关系。事务型数据库通常采用实体关系模型(ER 模型)进行逻辑设计。但是，ER 模型中各实体之间关系是对等的，不适用于数据仓库[3-4]。目前，数据仓库基本上使用维度模型(dimensional model)进行逻辑设计。在维度模型中，事实表(fact table)和维表(dimension table)是 2 种常用的数据库对象类型。事实表是一个大表，通常包含变化较快的、数值型的事实数据和指向维表的外键。而维表包含数据仓库中相对静态的文本型、描述性的数据。

本文利用星型维度模型对全国数据仓库进行建模，如图 2 所示。事实表中的数值数据为空气污染指数、各类污染物的浓度值和分指数，其他字段为指向各个维表的外键。按照对空气质量数据的分析角度，从时间、空间和属性 3 个方面设计了时间维表、城市维表、污染程度等级维表、空气质量级别维表、污染物类型维表等 5 个维表。从图 2 可以看出，对空气质量数据的任何查询都涉及事实表和维表之间的联合查询，这与 ER 模型的查询方式是完全不同的。

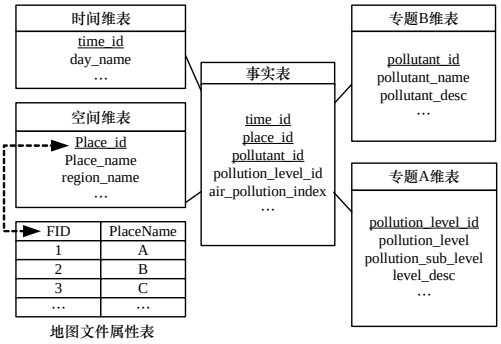

图 2 空气质量数据仓库的星型模型设计及扩展

为解决传统数据仓库无法灵活处理空间信息的缺陷，图 2 对星型模型进行了扩展，利用 ArcGIS9.0 的 ArcSDE 空间数据引擎将地图数据存储到 Oracle10g 数据库中，在空间维表中增加了一个 Place_id 字段以映射地图中标识空间图形的 FID 字段，将空间维的度量表现为空间对象的名称和指向空间对象的指针。这种 ID 值共享、图形与属性数据的动态关联技术，将空间信息引入到空间维中，实现了对空间数据的快速调用，能够方便地在空间维上对空气质量数据进行复杂分析和可视化表达。

3 空气质量数据仓库的物理实现

数据仓库的物理实现是将逻辑设计转化为物理数据库结构的描述，主要包括源数据的 ETL 过程和采用多种技术来提高查询性能，并使数据仓库易于维护。

3.1 源数据的 ETL

数据源主要包括 86 个城市每天的空气质量监测数据、中国《环境空气质量标准》、污染物及其危害的说明文件，以及全国行政区图和监测点空间分布图。

源数据在进入数据仓库之前，要进行提取、转化和装载(ETL)等操作来保证数据质量。空气质量监测原始数据存在记录重复、某些字段数据缺失、日期和地点违反唯一性约束、时间数据的格式不一致等诸多问题，通过对存在问题数据的筛查和不合格数据的后处理，保证了数据的一致性。

3.2 分区技术

事实表是数据仓库中最大的表，因此，必须考虑海量数据对存储和查询性能的影响。作为优化数据存储的重要方法，Oracle 的分区技术按照特定方式对大表进行逻辑划分，将数据部署到几个相对较小的分区段中。当访问分区表时，可以仅仅访问其中某个分区段，而不是整个表的所有数据。由于人们通常基于时间按照日、周、月、季度、年对空气质量对一个或多个城市的空气质量状况进行查询、比较和分析，因此本文对事实表按照年度进行分区。这样，如果要对某一年度的空气质量进行分析，只需要访问该年度所在的子分区，而无须查询其他年度的数据，因此，明显减少了访问的数据量，极大地提高了查询性能。

3.3 物化视图技术

数据仓库中包含了每天监测到的大量细节数据，然而人们在使用数据仓库时，通常需要统计查询和分析。如果每次查询都访问细节数据，则必然会降低查询效率。因此，大多数数据仓库都存储了多种粒度等级的数据。在 Oracle 数据仓库中，粗粒度的概要数据通过物化视图(materialized view)来实现。通过预先对聚集运算和连接操作进行计算，物化视图能够极大地提高查询分析速度。对于空气质量数据仓库而言，用户通常会查询以下一些问题：

(1)在空间上，按照城市、所属经济区来统计和分析各地的空气质量状况，如最大值、最小值、平均值、发生频率等。

(2)在时间上，按照天、周、月、季度和年统计不同时间段的空气质量状况。

(3)在内容上，关注空气质量的级别、首要污染物、污染物等级的时间发生次数和频率、空间发生次数和频率，并进行同期、异地的比较分析。

因此，设计时应该从这些问题出发来设计合适的物化视图。下列代码定义了物化视图 air_monthly_mv，用于计算各城市每月的最高(max)、最低(min)及平均(avg)空气污染指数。

```
create materialized view air_monthly_mv
build immediate refresh fast
enable query rewrite as
select t.month_desc, c.place_name,
   count(*), max(air_pollution_index), min(air_pollution_index),
   count(air_pollution_index), avg(air_pollution_index)
from air_china f, time t, cities c
where f.time_id = t.time_id and f.place_id = c.place_id
group by c.place_name, t.month_desc;
```

3.4 查询重写技术

查询重写(query rewrite)技术使得当用户对细节表进行统计查询时，Oracle 会自动将该查询重写到合适的物化视图上，即使用合适的物化视图(而不是细节表)来回答用户的查询。例如，在创建物化视图 air_monthly_mv 后，当用户使用如下语句访问细节表 air_china 来查询各城市每月的平均空气污染指数时，Oracle 会自动使用 air_monthly_mv，而不是 air_china 来回答用户的查询。

```
select t.month_desc, c.place_name, avg(air_pollution_index)
from air_china f, time t, cities c
where f.time_id = t.time_id and f.place_id = c.place_id
group by c.place_name, t.month_desc;
```

由于物化视图已经对统计查询进行了预先计算，因此查询重写将显著减少查询的响应时间。通常，随着数据数量的增加，物化视图和查询重写技术可以几十甚至几百倍地提高查询速度。同时，查询重写技术使得存储在数据仓库内部的物化视图对于用户或数据库应用程序是透明的。

4 空气质量数据仓库系统的 Web 空间 OLAP 系统

传统的 OLAP 服务大多基于关系模型和多维数据模型，无法做到对空间数据的几何图形进行全面分析，存在较大缺陷。而地理信息系统(GIS)技术专门面向空间数据的业务系统，擅长空间现象的表达和分析操作。因此，本文将 Web, GIS 和 OLAP 3 种技术相结合，设计了中国城市空气质量 Web Spatial OLAP 系统。该系统不仅能够通过 Web 方式满足各类用户多种查询和分析的需求，而且能够利用空间与属性数据的动态关联技术，访问监测数据与地图数据，实现专题地图制图功能，弥补传统数据仓库和 OLAP 工具在这方面的不足。

4.1 系统总体框架

中国城市空气质量 Web Spatial OLAP 系统由底部的数据源层、中间的数据驱动层和面向使用者的应用层等 3 层构成，如图 3 所示。

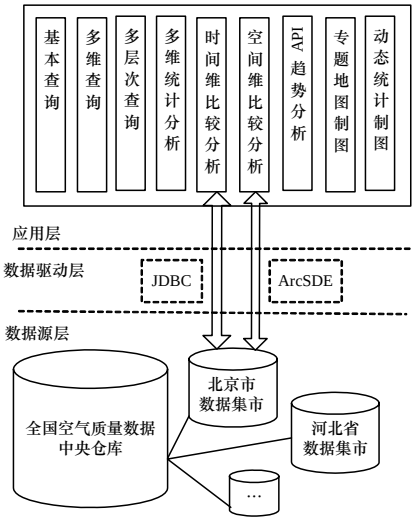

图 3 空气质量数据仓库 Web SOLAP 的系统框架

客户端通过 Web 界面，利用 JDBC 和 ArcSDE 访问中央仓库和数据集市的空气质量监测数据与地图数据，通过基本查询、多维查询、多粒度查询、多维统计分析、时间维比较分析、空间维比较分析、API 趋势分析、专题地图制图、动态统计制图等功能，以报表、多种形式统计图、专题地图等形式，向使用者提供有关我国城市空气质量的各类查询和分析结果，包括实报数值产品、统计分析产品、多维分析产品等。

4.2 系统部分功能介绍

本节以北京市空气质量数据集市为例，介绍 Web Spatial OLAP 分析系统的部分常用操作界面。

(1)基本查询

基本信息查询界面如图 4 所示，允许在任一时间段、任一(或全部)监测点、多个污染程度等级、多个空气质量级别、任意污染物的查询条件下，查询相应的空气质量状况，同时能够对查询结果按照多个维度排序。

图 4 基本信息查询界面

(2)多维查询和统计分析

多维查询和统计分析可以选择日、周、月、季度、年等 5 种时间粒度的任意一个时间段，以及监测站、区县、城市功能分区等 3 种空间粒度，查询空气质量等级、污染程度等级、污染物等级的统计数据。此外，能够用线状、二维柱状、三维柱状和三维饼图等多种形式对统计结果进行统计制图。

(3)专题地图制图服务

专题地图制图服务基于属性数据和空间数据动态关联技术，通过 JDBC 和 ArcSDE 访问数据仓库中空气质量监测数据与地图数据，利用 ArcIMS 的地图发布功能[5]，以 Web 形式提供任意时间北京市各监测站和各区县的空气质量状况的专题地图制图服务。例如，图 5 是根据 2006 年 12 月 14 日的北京市空气质量状况数据动态生成的专题地图。

图 5 北京市空气质量的专题地图界面

5 结束语

目前我国约有 200 个城市建立了空气质量监测体系，对几种主要污染物进行长期监测和分析，积累了大量的数据资源。本文以 2000 年至 2007 年我国主要城市空气质量数据为例，设计并实现了空气质量数据仓库，开发了一个 Web Spatial OLAP 系统，使用户能够以多粒度、多主题、多时期、多种显示模式(地图、表格、图形)的方式对空气质量监测数据进行访问、分析和可视化表达。通过利用空间与属性数据的动态关联技术，访问数据仓库中的空气质量监测数据与地图数据，有效解决了传统数据仓库和 OLAP 无法实现对空间数据的查询、分析和展现的缺点。今后将进一步加入气象数据、

（下转第 240 页）